MA334-Assignment

Data Analysis on the Biodiversity Measures

Reg no: 2211543

Introduction

Biodiversity, the variety of life on Earth, is an essential part of our planet's ecosystem. It is a measure of variability of species, genetics and the level of ecosystem. In recent times, the biodiversity has been degraded to its lowest extent as a result of recent environmental changes which led to mass extinctions. We have been provided with the data of biodiversity measures of *different taxonomic groups* in various dominant land class and time periods. The data here are based on the 11 taxonomic groups and I'll be performing analysis on selective 7 taxonomic groups which will be referred as BD7 with BD11. It does contains data in two different time periods such as Y70(from 1970) and Y00(from 2000). The data varies with location in all over Britain which has been segmented by grid squares measuring 100km across and also with different dominant land class accordingly. The segments are based on the *UK National Grid* in which the vertical lines are called Eastings and the horizontal lines are Northings. Their values tend to increase if you travel across east and north respectively. The mean of all taxonomic group has also provided, called **ecologicalStatus**.

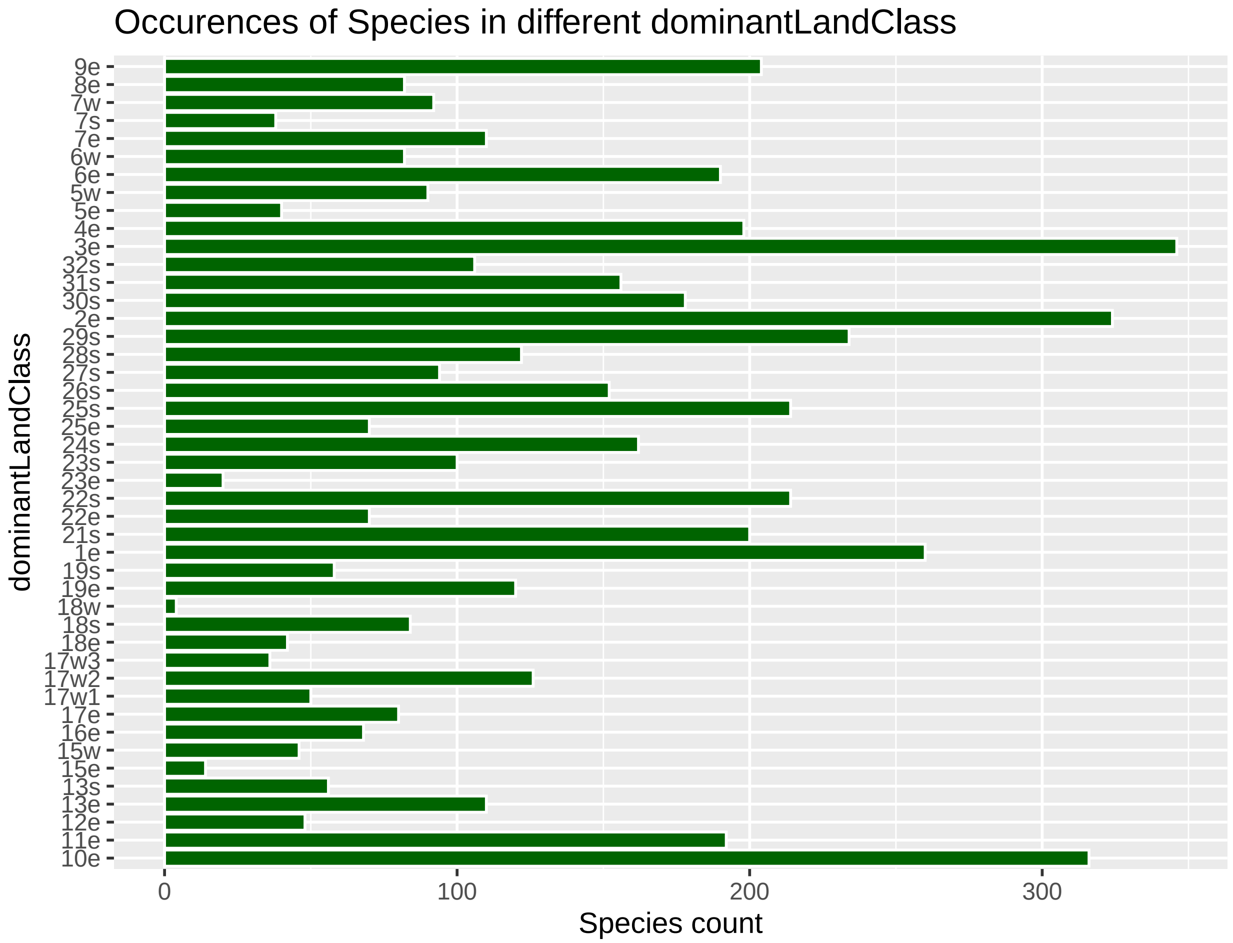
I'll be analyzing the data on biodiversity measure of the selected 7 taxonomic groups i.e, Bees, Bryophytes, Butterflies, Carabids, Hover flies, Grasshoppers Crickets, Vascular plants with all other taxonomic groups in different time periods and in various location. I'll also be performing some statistics on the data to provide insights into patterns and trends in biodiversity.

Main objectives:

1. Hypothesis Test
2. Simple Linear Regression
3. Multiple Linear Regression
4. Open analysis on the data

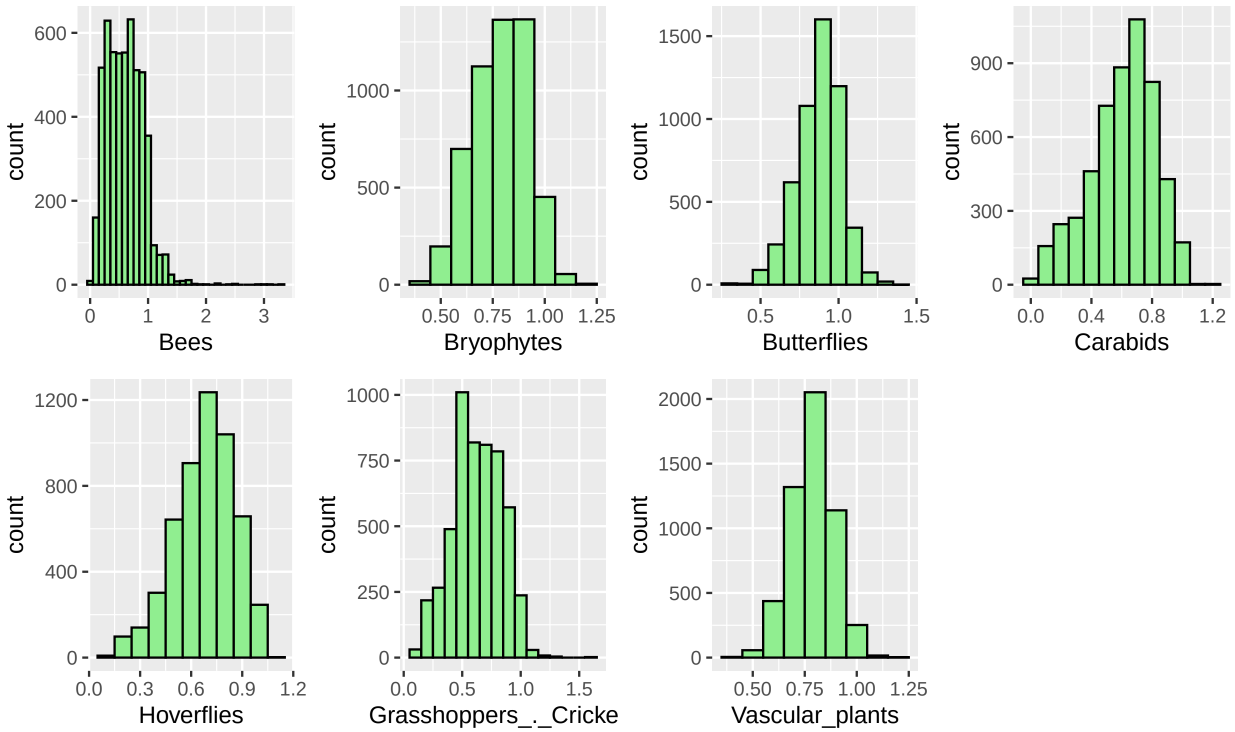
Data Exploration

The data consists of biodiversity measures for all taxonomic groups over two different time periods and across multiple locations. The *ecologicalStatus* is the mean of all the taxonomic groups' biodiversity measures and the *Northing*, *Easting* values are the coordinates used to accurately locate them on the map. The *dominantLandClass* is the names of the 45 different land classes in and around the UK. The following plot provides insights on the number of species in every dominantLandClass. The maximum number of species are from **3e** followed by **2e**, where the least are from *18w*.

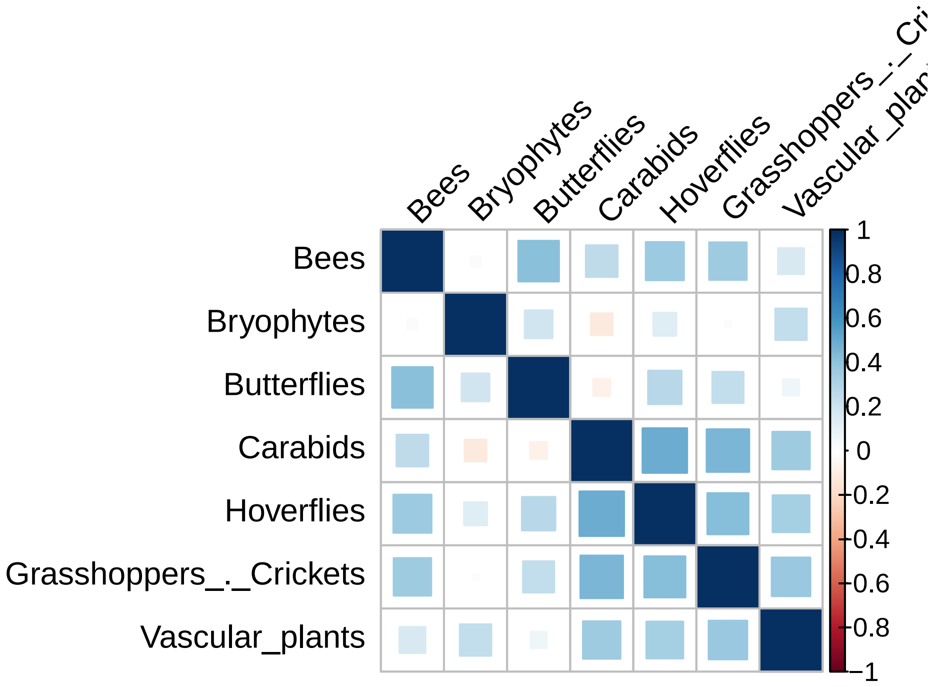


Here, we are looking at the distribution of each taxanomic groups and presenting their correlation to understand the relationship between the different selective taxonomic groups. From the histograms, we can interpret that some groups are **normally distributed** such as, butterflies, while others sit on a **skewed distribution**, like Bryophytes. We notice that the Bees and Butterflies possess **positive correlation** from the correlation matrix, whereas carabits are **negatively correlated**.

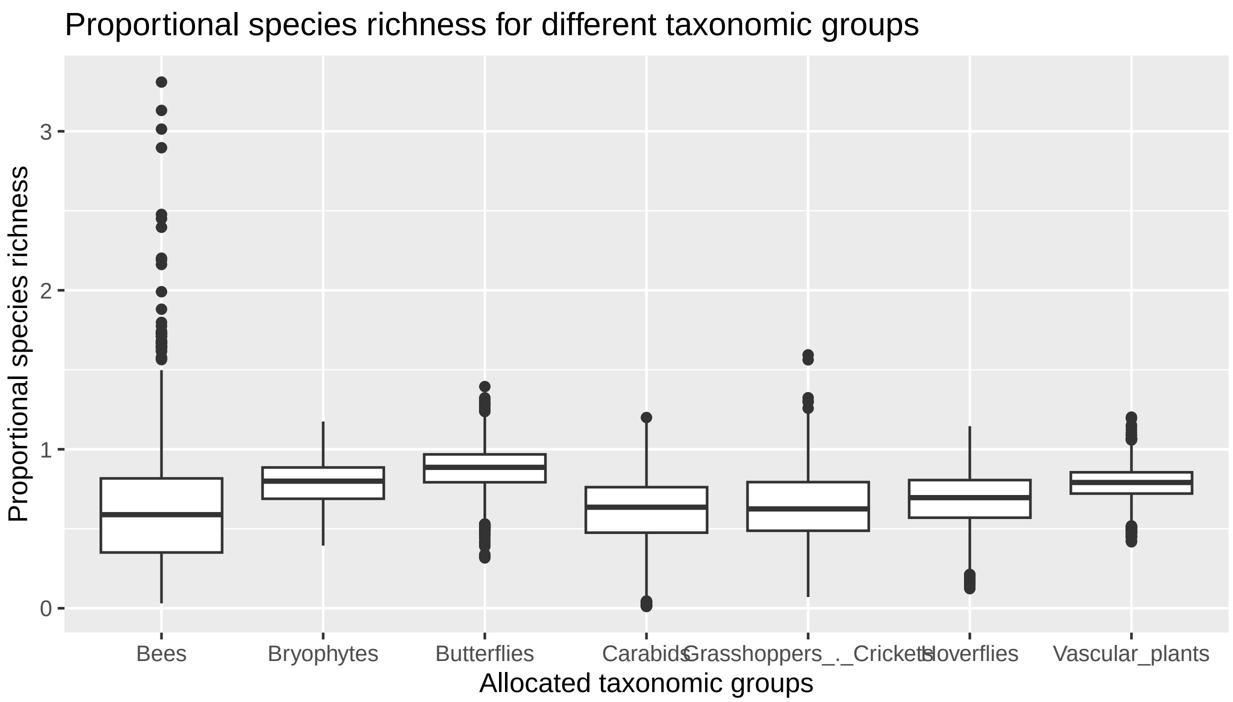
Distribution of the allocated taxonomic group



Correlation of 7 taxonomic groups

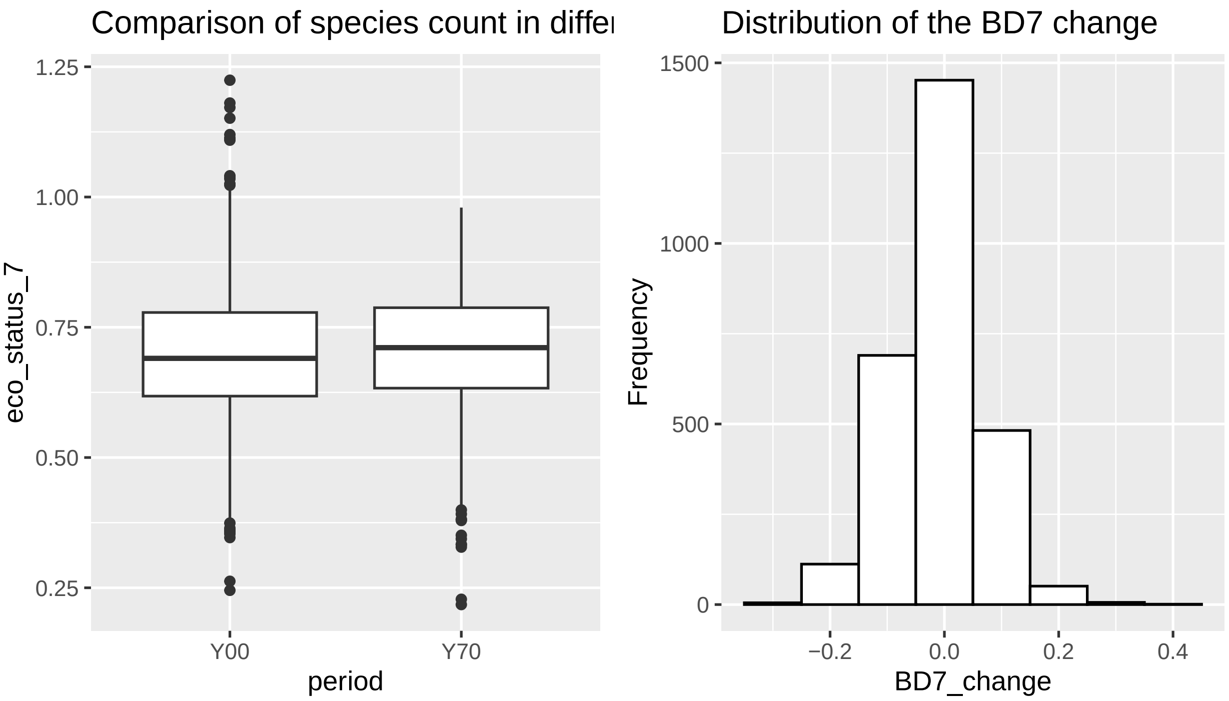


I have split all the 7 taxonomic groups accordingly and found the correlation between the location and the allocated taxonomic groups. As per the box plot, we found that the *Bees* are the dominating **taxonomic group** in terms of the proportional species richness among other groups followed by the grasshoppers. The amount of vascular_plants is too low, which pushes it down to the least variety of biodiversity.



Hypothesis Testing

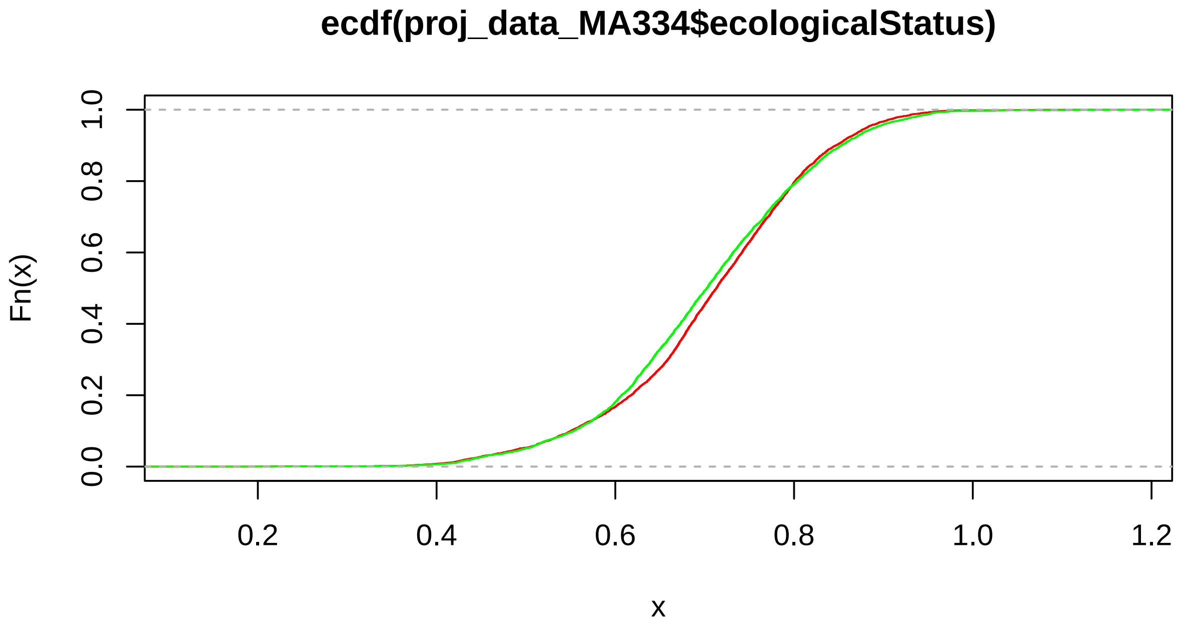
The following plot allow a visual comparison of *eco_status_7* over the two periods as well as the distribution of the change in *eco_status_7* in the two periods. The maximum number of outliers in the mean of the seven taxonomic group in the period **Y00** is significantly more than compared to the period **Y70**.



Performing a t.test with the *BD7_change*, which is nothing but the difference (Y00-Y70) of BD7 in two different time period. The frequency of BD7_change is equally distributed, which can be interpreted from the histogram. The null hypothesis of this test is that the population mean of the *BD7_change* is equal to 0. The alternate hypothesis is not equal to 0. The t-test here is used to check whether there is any difference between the mean of the *BD7_change* and the value 0. The t value, -8.2496 denotes that the no.of.standard errors in the sample mean is away from the hypothesized population mean. The df is 2798, and the p-value is 2.417e-16, which is less than the significance level. So the test suggests and supports the alternative hypothesis, as we have stronger evidence against the null hypothesis, which is that the mean of *BD7_change* is different from 0. The mean value lies somewhere between the 95 percent confidence interval, i.e., -0.01213813.

```
	One Sample t-test

data:  BD7_change
t = -8.2496, df = 2798, p-value = 2.417e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.015023199 -0.009253065
sample estimates:
  mean of x
-0.01213813
```

Providing with the plot of Empirical Cumulative distribution function (ecdf) between BD7 data and the BD11 data, which is the overall biodiversity measures of the taxonomic groups. Here the CDF, shows the probability that a random observation from the distribution is less than or equal to a given value, in which the green line represents the 7 taxonomic groups' mean and the red line denotes the overall mean. From the p-value *3.199e-08* observed from the Kolmogorov-Smirnov test, we can reject the null hypothesis and confirm that there are two different distributions.



```
    Asymptotic two-sample Kolmogorov-Smirnov test

data:  proj_data_MA334$eco_status_7 and proj_data_MA334$ecologicalStatus
D = 0.056627, p-value = 3.199e-08
alternative hypothesis: two-sided
```

The study analyzes the *descriptive statistics* of every biodiversity indices by taking their mean, standard deviation and the skewness. This shows the values tend to change before and after the time period.

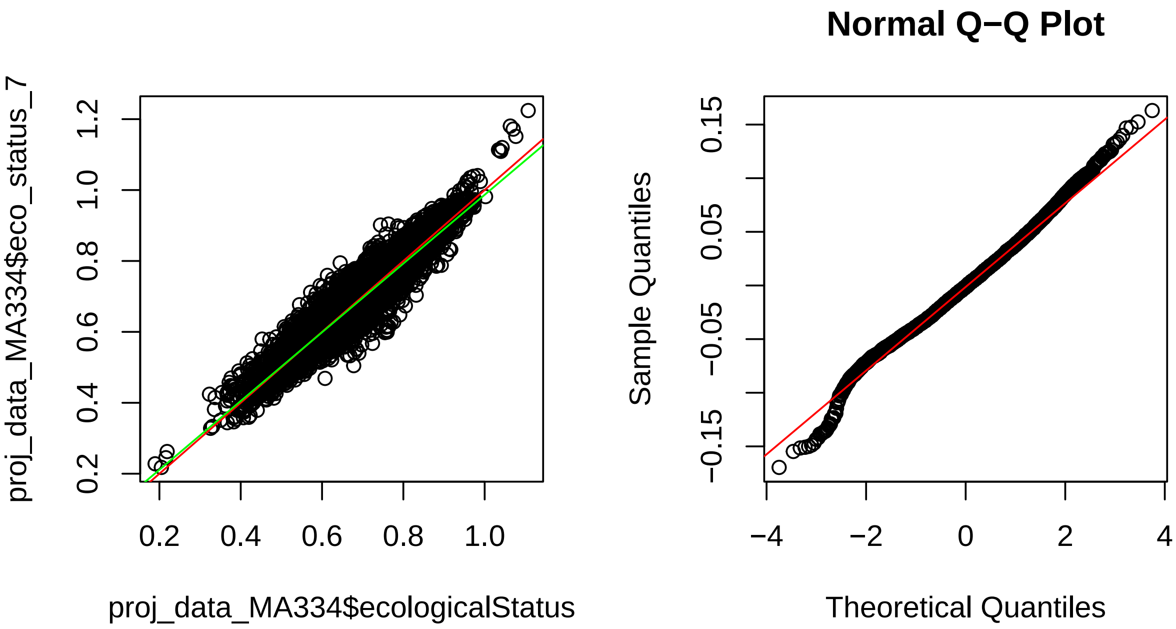
```
               taxi_group mean   sd skewness
1         Vascular_plants 0.78  0.1    -0.21
2              Bryophytes 0.79 0.13    -0.23
3             Butterflies 0.87 0.15    -0.47
4              Hoverflies 0.67 0.19    -0.52
5 Grasshoppers_._Crickets 0.62 0.21    -0.08
6                Carabids  0.6 0.22    -0.47
7                    Bees 0.59 0.31     0.97
```

Linear Regression

The linear regression for the *eco_status_7* with only Northing as a predictor has been carried out. We can observe from the values that the negative relationship between the mean of 7 taxonomic group and the northing, as the coefficient of northing is -1.808e-07 with the standard error of 5.043e-09. The northing value has 18.68% tendency on the variation of *eco_status_7*. The p-value indicates that the northing can significantly improve the fit of the model. The residuals has been plotted in the qq-plot, which shows a straight line which means that the residuals are normally distributed.

Another simple linear regression between the 7 allocated taxonomic groups with the overall taxonomic groups and displaying the linReg values with different time periods. The regression line(red) and the identity line(green), which represents the perfect fit, has been overlaid on the scatter plot. Then using the qqnorm() on the residuals will give us information on the assumption of normality for the residuals. The intercept values demonstrates a positive association between the mean of all the taxonomic groups and the mean of the selected 7 taxonomic groups in both the periods y00 and y70.

```
Call:
lm(formula = proj_data_MA334$eco_status_7 ~ proj_data_MA334$ecologicalStatus)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.169597 -0.027686 -0.001266  0.024937  0.163067 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      0.017305   0.003271    5.29 1.27e-07 ***
proj_data_MA334$ecologicalStatus 0.969322   0.004573  211.98  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.03945 on 5596 degrees of freedom
Multiple R-squared:  0.8893,    Adjusted R-squared:  0.8892 
F-statistic: 4.493e+04 on 1 and 5596 DF,  p-value: < 2.2e-16
```



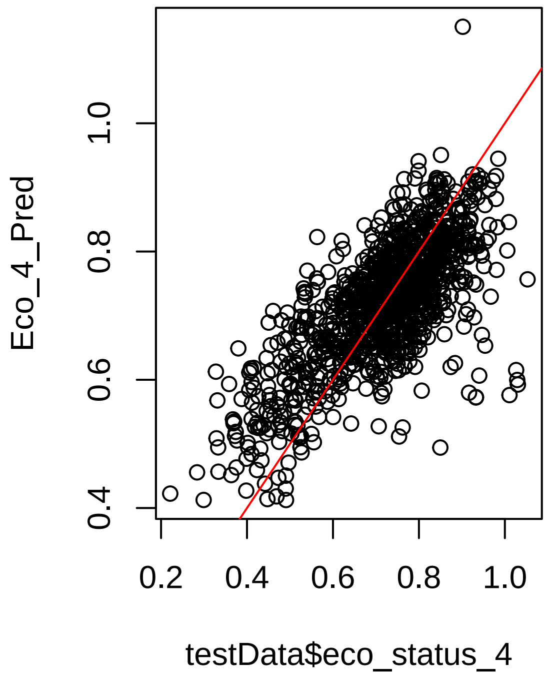
```
(Intercept) proj_data_MA334_Y70$ecologicalStatus
 0.03404394                           0.94542215

(Intercept) proj_data_MA334_Y00$ecologicalStatus
0.002484931                          0.990926155
```

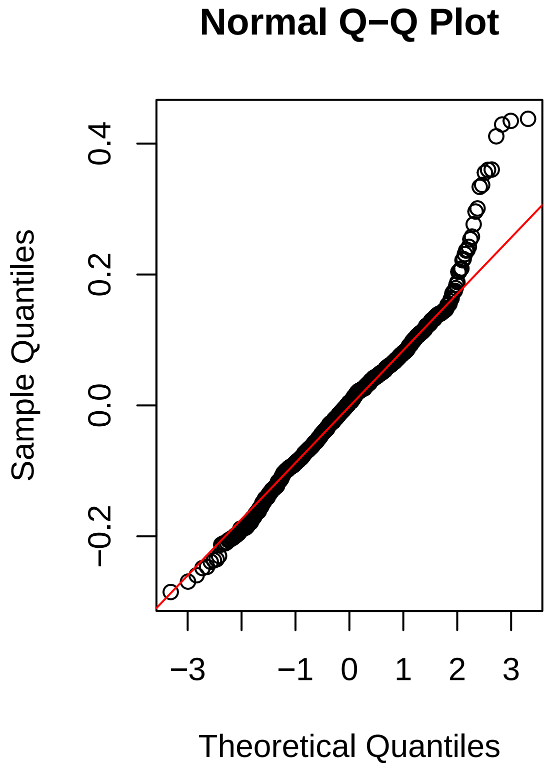
Multiple Regression

Here, I am performing *multiple regression* between the **BD7 groups** and **BD4 groups**. First the data has been split into training 80% and testing 20%, then its trained using the training data and all other predictors of BD4. Then the lm() function fits the multiple regression model and displays the output. With the model making predictions on the test dataset, the correlation between the actual values and the predicted values has been computed and the scatter plot has been presented with the line indicating the perfect predictions. Even the differences between the actual values and the predicted values has also been plotted to check for any patterns that may indicate any issues with the model fit. Finally a QQ plot with the differences has been presented to check for normality. From the *p-value*, we can say that the taxonomic groups except the grasshopper__crickets, have a strong positive relation with the ecologicalStatus. The model explained 54.5% of the variation in BD11 mean from the R-squared values. Overall the model is statically significant based on the p-values and the F-statistic values.

```
Call:
lm(formula = eco_status_4 ~ ., data = trainingData[c(eco_selected_names,
    "eco_status_4")], na.action = na.omit, y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33998 -0.05850  0.00319  0.05674  0.47041
```

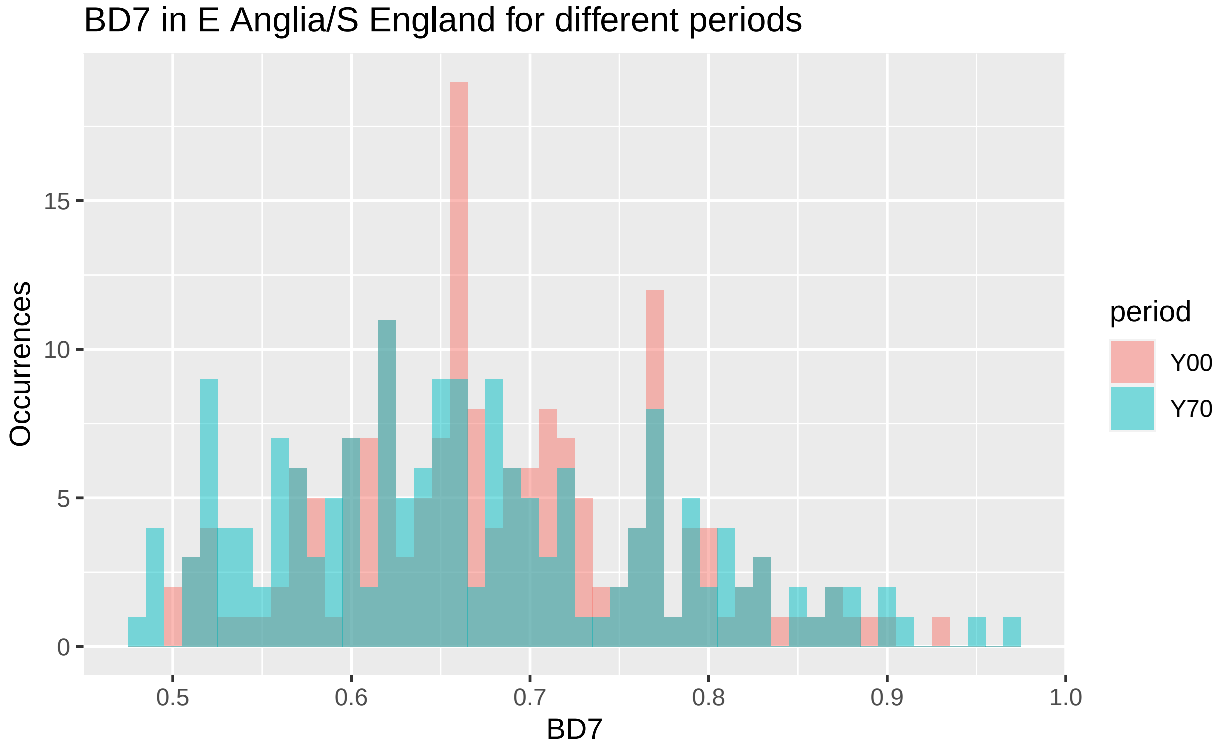
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               0.322141   0.014314  22.505   <2e-16 ***
Bees                      0.101426   0.005323  19.055   <2e-16 ***
Bryophytes               -0.221163   0.011336 -19.509   <2e-16 ***
Butterflies               0.171907   0.011601  14.818   <2e-16 ***
Carabids                  0.149285   0.008392  17.788   <2e-16 ***
Hoverflies                0.236911   0.009897  23.937   <2e-16 ***
Grasshoppers_._Crickets -0.016096   0.008171  -1.970   0.0489 *  
Vascular_plants           0.154101   0.015767   9.774   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.09267 on 4429 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.5517,    Adjusted R-squared:  0.551 
F-statistic: 778.7 on 7 and 4429 DF,  p-value: < 2.2e-16
```




Normal Q−Q Plot

Open Analysis

From the plot earlier in the report, which shows the number of species over the *dominantLandClass*, we came to know that the majority of species were from the class *‘3e’* i.e, **E Anglia/S England**. So examining the dominantLandClass, *‘3e’* with two different time periods. The following plot compares the BD7 values, which are the mean of the allocated taxonomic groups in East Anglia/South England between two time periods: Y00 & Y70. The histogram revealed that the BD7 values increased between the two periods, hitting the maximum in Y00 period, which describes that areas in Y00 with a good ecological status. Comparatively, the species count was lower in the areas with lowest ecological status in Y70. This highlights the importance of monitoring the biodiversity over period to gain insights and act accordingly to preserve it.



Conclusion

Based on the **overall analysis** on the biodiversity measures of the selective taxonomic groups *BD7* and with all other groups *BD11*, we interpret that the data doesn't shift drastically but we could identify some trends and able to understand the presence of biodiversity and its decline. Additionally we also examined the changes over the values with the *BD4 taxonomic groups*. We found the correlation significance between taxonomic groups, and some were positively correlated like bees, butterflies, while others are negatively correlated, such as carabids.

Finally, we found that the abundance of several taxonomic groups including bees, bryophytes, butterflies, etc, can be used as predictors of the ecological status of an area. We also found the amount of species differ over time period Y00 and Y70, which taught us the importance of monitoring the biodiversity measures to prevent the ecosystem from extinction. Our findings suggests there is a strong decline in the biodiversity and there should be measures taken in order to increase the conservation efforts to protect the biodiversity for a sustainable ecosystem in the future.

References

1. Wikipedia contributors. (2023). Biodiversity. In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Biodiversity

2. Fountain, T., et al. (2015). The impacts of climate warming on the properties of boreal forests in Alaska. Journal of Ecology, 103(2). https://doi.org/10.1111/1365-2664.12784